A Broader Impact

This paper aims to provide a fair federated learning algorithm to guarantee that the utilities of the trained agents are core-stable fair. We do not expect the work to have any ethics issues or negative social impact if it is correctly used. On the other hand, if our evaluation and theory is misused, there could be potential negative social impact. For instance, our fairness metrics cannot indicate other accuracy or loss utilities and people need to evaluate federated learning algorithms with different utility metrics, rather than only using our metrics. We expect that our work will provide a way to measure and achieve fairness for different federated learning paradigms.

B Missing proofs from Section 4.1

B.1 Proof of Lemma 4.1

Proof. We prove by contradiction. Assume that there exists a $\theta' \in P$, and a $S \subseteq [n]$, such that $(|S|/n) \cdot u_i(\theta') \geq u_i(\theta)$ for all $i \in S$ with at least one strict inequality. Then, we have $\frac{u_i(\theta')}{u_i(\theta)} \geq n/|S|$ for all $i \in S$ with at least one strict inequality, implying $\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta)} \geq \sum_{i\in S} \frac{u_i(\theta')}{u_i(\theta)} > n$. However, since $\theta \in \phi(\theta)$, we have $\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta)} \leq \sum_{i\in[n]} \frac{u_i(\theta)}{u_i(\theta)} = n$, which is a contradiction. □

B.2 Proof of Lemma 4.2

Proof. We need to show that for every sequence $(\theta)_i$ converging to θ, and $(\gamma)_i$ converging to γ, such that $\gamma_i \in \phi(\theta_i)$ for all i, we have $\gamma \in \phi(\theta)$. We prove this by contradiction. Let us assume otherwise, $\gamma \notin \phi(\theta)$. Let $\gamma' \in \phi(\theta)$ and let $\delta = \frac{\sum_{i\in[n]} u_i(\gamma')/u_i(\theta)}{\sum_{i\in[n]} u_i(\gamma)/u_i(\theta)} > 1$. We now make a technical claim about the utility functions of the agents.

Claim B.1. *For all $i \in [n]$, and x, y such that $||x - y||_2 \leq \beta$, we have*

1. *$|u_i(x) - u_i(y)| \leq h(\beta)$ where $h\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is continuous increasing function with $h(0) = 0$, and*
2. *for each $i \in [n]$, we have $u_i(y) \cdot h'(\beta)^{-1} \leq u_i(x) \leq u_i(y) \cdot h'(\beta)$, where $h'(\beta) = (1 + \frac{h(\beta)}{M\varepsilon})$ and $M = min_{i\in[n]} M_i$.*

Proof. Claim (1) follows immediately from the continuity of the utility functions.

For claim (2), we have

$$\begin{aligned} u_i(x) &\leq u_i(y) + h(\beta) \\ &\leq u_i(y) \cdot (1 + \frac{h(\beta)}{u_i(y)}) \\ &\leq u_i(y) \cdot (1 + \frac{h(\beta)}{M_i\varepsilon}) && (u_i(y) \geq M_i\varepsilon) \\ &\leq u_i(y) \cdot (1 + \frac{h(\beta)}{M\varepsilon}) && (M_i \geq M) \\ &\leq u_i(y) \cdot h'(\beta). \end{aligned}$$

In a similar way, we can prove that $u_i(y) \leq u_i(x) \cdot h'(\beta)$, which would then imply that $u_i(x) \geq u_i(y) \cdot (h'(\beta))^{-1}$. □

We choose a δ' such that $h'(\delta')^3 = (1 + \frac{h(\delta')}{M\varepsilon})^3 \ll \delta$. Such a δ' exists as $h()$ is a continuous increasing function with $h(0) = 0$, and $\delta > 1$. Since the sequences $(\theta)_i$ and $(\gamma)_i$ converges to θ and γ respectively, there exists a $n' \in \mathbb{N}$ such that for all $\ell \geq n'$, we have $||\gamma_\ell - \gamma||_2 < \delta'$ and

$||\theta_\ell - \theta||_2 < \delta'$. Now observe that

$$\begin{aligned}
\sum_{i\in[n]} \frac{u_i(\gamma')}{u_i(\theta_\ell)} &\geq h'(\delta')^{-1} \cdot \sum_{i\in[n]} \frac{u_i(\gamma')}{u_i(\theta)} && \text{(by Claim B.1)} \\
&= h'(\delta')^{-1} \cdot \delta \cdot \sum_{i\in[n]} \frac{u_i(\gamma)}{u_i(\theta)} && \text{(by definition of } \delta) \\
&\geq h'(\delta)^{-2} \cdot \delta \cdot \sum_{i\in[n]} \frac{u_i(\gamma_\ell)}{u_i(\theta)} && \text{(by Claim B.1)} \\
&\geq h'(\delta)^{-3} \cdot \delta \cdot \sum_{i\in[n]} \frac{u_i(\gamma_\ell)}{u_i(\theta_\ell)} && \text{(by Claim B.1)} \\
&> \sum_{i\in[n]} \frac{u_i(\gamma_\ell)}{u_i(\theta_\ell)} && \text{(as } \delta \gg h'(\delta')^3\text{).}
\end{aligned}$$

This shows that $\gamma_\ell \notin \phi(\theta_\ell)$, which is a contradiction. □

C Missing Proofs from Section 4.2

C.1 Proof of Theorem 2

Proof. We first show that for any other predictor $\theta' \in P$, we have $\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta^*)} \leq n$. Consider any other predictor $\theta' \in P$. Since P is convex, we have $(\nabla_\theta \mathcal{L}(\theta^*))^T(\theta' - \theta^*) < 0$. Now, observe that

$$\begin{aligned}
\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta^*)} - n &= \sum_{i\in[n]} \frac{u_i(\theta') - u_i(\theta^*)}{u_i(\theta)} \\
&\leq \sum_{i\in[n]} \frac{(\nabla u_i(\theta^*))^\intercal(\theta' - \theta^*)}{u_i(\theta^*)} && \text{(from concavity of } u_i()\text{)} \\
&= \sum_{i\in[n]} \sum_{j\in[d]} \Big(\frac{\partial u_i(\theta^*)}{\partial \theta_j} \cdot (\theta'_j - \theta^*_j) \cdot \frac{1}{u_i(\theta^*)}\Big) \\
&= \sum_{i\in[n]} \frac{1}{u_i(\theta^*)} \cdot \sum_{j\in[d]} \Big(\frac{\partial u_i(\theta^*)}{\partial \theta_j} \cdot (\theta'_j - \theta^*_j)\Big) \\
&= \sum_{j\in[d]} (\theta'_j - \theta^*_j) \cdot \sum_{i\in[n]} \Big(\frac{1}{u_i(\theta^*)} \cdot \frac{\partial u_i(\theta^*)}{\partial \theta_j}\Big) \\
&= (\nabla_\theta \mathcal{L}(\theta^*))^\intercal(\theta' - \theta^*) < 0.
\end{aligned}$$

Now if θ^* is not core-stable, then there exists an $S \subseteq [n]$ and $\theta' \in P$, such that $u_i(\theta') \geq n/|S| u_i(\theta^*)$ for all $i \in S$ with at least one strict inequality, then we have $\sum_{i\in[n]} \frac{u_i(\theta)}{u_i(\theta')} \geq \sum_{i\in S} \frac{u_i(\theta)}{u_i(\theta')} > n/|S| \cdot |S| = n$, which is a contradiction. □

D Algorithm CoreFed

Here we present the full description of CoreFed in Algorithm 1.

E Missing Proofs from Section 4.4

E.1 Proof of Theorem 3

Proof. For any θ' such that $||\theta - \theta'||_2 \leq d$, according to the definition of β-smooth, we have

$$u_i(\theta') \leq u_i(\theta) + \nabla_\theta u_i(\theta)^\intercal(\theta' - \theta) + \frac{\beta}{2}||\theta' - \theta||_2^2.$$

Algorithm 1: CoreFed Distributed Training Protocol.

Input: Number of clients K, number of rounds T, epochs E, learning rate η
Output: Model weights θ^T

for $t = 0, 1, \cdots, T-1$ **do**

Server selects a subset of K devices S_t;

Server sends weights θ^t to all selected devices;

Each select device $s \in S_t$ updates θ^t for E epochs of SGD with learning rate η to obtain new weights $\bar{\theta}_s^t$;

Each select device $s \in S_t$ computes

$$\Delta\theta_s^t = \bar{\theta}_s^t - \theta^t,$$
$$\mathcal{L}_s^t = \frac{1}{|\mathcal{D}_s|} \sum_{i=1}^{|\mathcal{D}_s|} \ell(f_{\theta^t}(x_s^{(i)}), y_s^{(i)})$$

where $\mathcal{D}_s = \{(x_s^{(i)}, y_s^{(i)}) : 1 \le i \le |\mathcal{D}_s|\}$ is the training dataset on device s;

Each selected device $s \in S_t$ sends $\Delta\theta_s$ and $\mathcal{L}_s$ back to the server;

Server updates θ^{t+1} following

$$\theta^{t+1} \leftarrow \theta^t + \frac{1}{|S_t|} \sum_{s \in S_t} \frac{\Delta\theta_s^t}{M_s - \mathcal{L}_s^t}.$$

end

Then we observe that

$$\begin{aligned}
&\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta)} - kn \\
=&\sum_{i\in[n]} \frac{u_i(\theta') - u_i(\theta)}{u_i(\theta)} - (k-1)n \\
\le&\sum_{i\in[n]} \frac{\nabla_\theta u_i(\theta)^\intercal(\theta'-\theta) + \frac{\beta}{2}||\theta'-\theta||_2^2}{u_i(\theta)} - (k-1)n \\
=&(\nabla\mathcal{L}(\theta))^\intercal(\theta'-\theta) + \sum_{i\in[n]} \frac{\beta}{2u_i(\theta)}||\theta'-\theta||_2^2 - (k-1)n \\
\le&\epsilon||\theta'-\theta||_2 + \sum_{i\in[n]} \frac{\beta}{2u_i(\theta)}||\theta'-\theta||_2^2 - (k-1)n.
\end{aligned}$$

By plugging in the RHS of Equation (4), we observe that when $||\theta'-\theta||_2 < d$, $\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta)} - kn < 0$.

On the other hand, suppose for any $S \subseteq [n]$, if for all $i \in S$ we have $\frac{|S|}{kn} u_i(\theta') \ge u_i(\theta)$, then

$$\sum_{i\in[n]} \frac{u_i(\theta')}{u_i(\theta)} = \sum_{i\in S} \frac{u_i(\theta')}{u_i(\theta)} + \sum_{i\in[n]\setminus S} \frac{u_i(\theta')}{u_i(\theta)} \ge \sum_{i\in S} \frac{u_i(\theta')}{u_i(\theta)} \ge \sum_{i\in S} \frac{kn}{|S|} \ge kn \tag{5}$$

which contradicts the above result and concludes the proof. □